High Educ (2014) 67:273–288
DOI 10.1007/s10734-013-9649-1

The futility of attempting to codify academic achievement standards

D. Royce Sadler

Published online: 5 July 2013

Abstract Internationally, attempts at developing explicit descriptions of academic achievement standards have been steadily intensifying. The aim has been to capture the essence of the standards in words, symbols or diagrams (collectively referred to as codifications) so that standards can be: set and maintained at appropriate levels; made broadly comparable in different specified contexts; and generally shared and understood better by assessors, academic program directors, students, employers, quality assurance agencies and the public at large. The scale of this practice ranges from rubrics for single assessment tasks to national standards statements used as academic performance benchmarks for graduates from academic programs. A critical analysis shows that the underlying assumptions of this process are fundamentally flawed. Codifications are inherently incapable of meeting the requirements because key terms lack the necessary attributes. A fundamentally different material form of representation is therefore necessary if the original intentions are to be realised.

Keywords Academic standards · Achievement · Quality assurance · Rubrics · Codification · Grading · Comparability

Introduction

In many higher education contexts, decisions on assessment and grading and in some cases the content of courses are devolved to individual academics, small teams or program directors. A growing practice has been to develop explicit descriptions of expected standards so they can be used by students (as producers) and academic appraisers. The basic idea is simple. Specifications of desired outcomes and standards are compact, reproducible

D. Royce Sadler (✉)
Teaching and Educational Development Institute, The University of Queensland, Learning Innovation Building, St Lucia, QLD 4072, Australia
e-mail: d.sadler@uq.edu.au

and portable. If all relevant parties work to the same set of specifications, the belief is that appropriate levels of consistency and comparability will result. This article is primarily a critique of that hypothesis. It is not about setting standards, although that is a crucial topic in its own right.

The specifications can be as simple as a teacher developing and using a rubric to improve transparency and accountability about what is expected from students when they produce complex responses to an assessment task. Students can shape their work accordingly, and markers can mark to the same rules. At the other extreme is a quality assurance agency with legislative authority to ensure that, as far as possible, all graduates from degree programs satisfy minimum requirements of knowledge and skills in various disciplines and professions. (Instead of a quality assurance agency, a voluntary consortium of institutions may pursue the same quality-related goal). General practice is to set out degree expectations in the form of learning outcomes, of which one subset is to be attained by all students graduating from all degree programs in all institutions, regardless of field. This subset usually goes under a label such as 'graduate' or 'generic' followed by 'outcomes', 'skills', 'capabilities', 'competencies' or 'attributes'. Typical generic outcomes are: critical analysis; problem solving; locating, evaluating and using information; originality, initiative and creativity; and effective communication. Other expected outcomes are specific to particular disciplines, fields and professions. Broad comparability of achievements of graduates from different higher education institutions within a given jurisdiction would also facilitate student mobility with credit transfer. Extending comparability across provincial or national borders would facilitate mutual recognition of professional qualifications. This commitment to assuring the overall quality of higher education has resulted in significant investment in developing comparability procedures.

The initiative along these lines with the widest international reach is the Tuning project, which is part of the well-documented Bologna Process. Although Tuning began in the European Union, the aims and methodology are also being pursued in some non-EU European countries, the USA, Canada, Latin America and Africa. Parallel initiatives, not under the Tuning banner, are in progress in the UK with its 'subject benchmark statements' and in Australia with its 'threshold learning outcomes'. In large-scale implementations, wide-ranging consultative processes in different disciplines, fields and professions (including sciences, technologies and mathematics) seek to identify common ground without standardisation of curriculum, teaching approaches or testing procedures. The goals of comparability and equivalence are not intended to produce uniformity. Extensive consultation should also, as a side benefit, foster a sense of ownership of the final result. However, the analysis in this article will show why consensus on the wording of the outcome statements does not necessarily result in consensus on underlying achievement standards.

Terminology

In this article, *codification* and *coding* have distinct meanings. Explicit descriptions or statements composed of words, diagrams or symbols are referred to as *codifications*. Laws, policies, game rules, regulations and instruction manuals are all examples of codifications. To a large extent, so are the scientific journal articles that report discoveries which can be replicated by others. Ideally, codifications will 'serve to "reconstitute" knowledge at a later time, in a different place, or by a different group of individuals' (Amin and Cohendet 2004 p. 21). In the context of assessment and grading, codifications include: rubrics,

criteria-standards matrices, marking guides, scoring schemes, grade descriptors, minimum (threshold) standards, subject or discipline benchmark statements, and graduate attributes. These are regarded as primary tools for communicating, transferring and sharing 'standards knowledge' among learners, academics, accreditation agencies, professional bodies and employers.

At the course level, standards codifications may include descriptions which differentiate two or more levels of achievement. Each level is given a label, which is its *code*, typically an alphanumerical symbol or word phrase such as Pass/Fail; A, B, C, D, F; Distinction, Merit, Credit, Pass, Fail; or 5, 4, 3, 2, 1, 0. The achievement information represented in codes is highly condensed and needs contextualisation for meaningful interpretation. *Codification* is the act of developing explicit descriptions of standards, while *coding* is the act of representing a level of achievement by a mark, rating or grade.

Another distinction is important. In ordinary usage, *criteria* and *standards* are often used interchangeably and the context makes the meaning clear. The problem in many discussions on educational achievement is that the meanings of these two terms slide around and lead to confusion. In this article, keeping them distinct serves a useful purpose. A *criterion* is a property or quality used in appraising student responses to assessment tasks, whereas a *standard* is a minimum achievement level used as a reference point when judging the quality of a student's work so the appropriate code can be assigned to it (Sadler 1985, 1987). Standards are underpinned by criteria (as qualities) but criteria as qualities can make sense without reference to standards.

Certain concepts and terminology used in the analysis resemble those used in semiotics, but no appeal is made to semiotic theory as such. The term *specifier* refers to a word or short phrase used as the label for something; the term *specified* refers to the thing so identified. The specified may be a property, a physical substance, an amount, a class, a concept, or something else altogether, abstract or concrete. Particular specifiers and specifieds take on not only meaning but also practical relevance as a duality embedded in actual contexts. Later on, the term *referent* is also used to denote a specific instance or example of something referred to by a specifier.

Structure of the argument

The argument about to be made is that for a codification to be adequate for the carriage of academic achievement standards, certain properties must characterise its principal elements. These principal elements are qualities (or characteristics) of student works or performances and amounts or levels of those qualities which must be present for a particular standard to be attained. Qualities which refer to achievement or performance are not directly observable or measurable. Levels of attainment or achievement are compounded from inferences based on evidence such as student responses to assessment tasks. The argument has two branches. First, achievement is not a physical variable but a concept which has fuzzy boundaries. Second, the words used to designate amounts are elastic in their interpretation. The meanings for the principal elements, qualities and amounts alike, are inherently context dependent. They are not and cannot be standardised. Since contexts differ, a stand-alone codification cannot be interpreted in a unique way by different people in different contexts at different times. Codifications therefore cannot 'hold' standards by serving as stable reference points for judging and reporting different levels of student achievement.

This conclusion is independent of whether the description is a rubric for a single assessment task or a benchmark standard for an entire degree program. It is also

independent of the specificity of the descriptors—coarse, fine or superfine. Codifications may well indicate dimensions or aspects of importance, but are theoretically incapable of adequately representing standards. If the goal of 'assuring' standards in a variety of contexts without implementing uniform curriculum and assessment practices is considered worthwhile, it has to be approached through other means.

The argument is developed by analysing the attributes of the word elements in three examples of educational achievement 'standards'. In order to clarify the essential character of the principal elements, a codified standard from an entirely different field is discussed. This external codification is able to hold and convey the intended standard, and provides a clean, stark contrast to codified standards in education. Some structural similarities exist, but the differences are significant enough to establish that the two types of codifications are situated in distinct categories, and differ not simply in degree. However, the external codification cannot serve as a model for emulation in the academic context, or be adapted to it.

Typical higher education 'standards' codifications

The following three examples of codified achievement standards all involve multiple criteria and are adaptations from real cases. Subject or discipline identifiers have been removed and grammatical forms within each example have been made consistent. The specifiers of interest refer to qualitative and quantitative characteristics. The qualitative specifiers identify the criteria—the qualities, properties, characteristics or features of interest. The quantitative specifiers are the amounts or levels required to qualify for a particular classification or gradation.

Table 1 shows a format applied to threshold (minimum) capabilities expected of all graduates and approved by a quality assurance agency. A complementary table could show somewhat higher overall performance standards that are expected of the *majority* of (rather than *all*) graduates, making two fixed reference levels. The standards in Table 2 are similar in format, with the four levels matching an institutional grade scale. The descriptors are intended to refer to achievement levels reached by the end of each course. Table 3 sets out a typical rubric for written or other extended responses to assessment tasks. Some Objective Structured Clinical Examinations (OSCEs) make use of similar scoring rubrics. The rubric in Table 3 has three criteria cross-tabulated with four levels of performance ('standards') for each criterion as spelled out by the text in the cells. Scoring involves deciding which text cell on each criterion best fits the student work under scrutiny. An overall score may be arrived at by inspection of the patterns of cells selected or, if numbers are assigned to the cells, by weighting and adding the numbers.

In Table 3, the main criteria are supplemented by more criteria embedded in the text cells: *accuracy*, *integration*, *logical development* and *support for assertions*. Potentially, each could be elaborated further. *Integration*, for example, could be expanded as 'how well a work comes together as a whole rather than as separate pieces, how the different aspects of the work are linked and smoothly combined into a unified entirety'. Although containing redundancies, this filling out of the meaning may be useful to a student or a marker. But observe that this expansion of *integration* shows how it intersects conceptually with other stated criteria, specifically *consistency of focus*, *relatedness of key ideas*, *support for assertions* and *progression of ideas*. Weaknesses in any of the latter actively reduce the level of *integration*. This phenomenon is common in achievement codifications. Criteria which may appear to be distinct in the abstract are often found to overlap, and occasionally

Table 1 Typical graduate learning outcomes for an undergraduate degree

Each graduate is expected to demonstrate:
Comprehensive understanding and systematic awareness of the basic sciences fundamental to both the discipline and professional practice;
Ability to locate, evaluate, summarise and apply primary and secondary sources of information;
High-level proficiency in a range of scientific and clinical techniques, including collection, analysis and critical interpretation of data;
Ability to apply principles and techniques to solve problems in routine and novel situations;
Willingness to take personal responsibility and exercise initiative in unpredictable and complex situations;
Capacity to make informed and reliable judgements, in a professional manner, including decisions with incomplete data;
Skill in communicating effectively the scientific aspects of their work to both specialist and non-specialist audiences; and
Competence in constructing reasoned arguments to support their actions and positions on the ethical and social impact of research in the discipline and professional practice.

Table 2 'Standards' descriptors for grades A to D in a course or module

Grades	Standards descriptors
A	Deep, broad knowledge of the course content; strong evidence of highly original thought; excellent analytical and critical abilities as well as a thorough grasp of the topic from both background reading and independent research; exemplary organisational, rhetorical and presentational skills in oral and written modes
B	Generally adequate knowledge and understanding of most of the topics, with higher levels of mastery in some areas; moderate level of critical and abstract thinking; some originality and independence in evaluating and organising information; generally competent in communicating through written and oral presentations
C	Broad knowledge and comprehension of much of the primary course content; some ability to apply knowledge in routine situations, but markedly less ability in novel contexts; analytic and critical ability limited; written communication skills mostly sound except for extended or complex material
D	Narrow perspective and marginal knowledge of the field; very limited capacity to confront non-routine issues or problems; virtually no ability to locate and assemble information independently; notable lack of rigour in analytic thinking; oral and written presentations often inaccurate, lacking in logic and inclusive of irrelevant material

even to interfere, with other criteria when an attempt is made to apply them meticulously. The various properties are not therefore mutually independent but, whether taken together or in clusters, refer to compounded qualities and meanings.

Difficulties of a somewhat different type arise through the particular criteria that are explicitly or implicitly embodied in a rubric. In the process of rubric construction, choices among criteria have to be made. Some criteria are included, others not. It is immaterial whether this occurs by deliberate decision, by not thinking of some criteria at the time, or by not even being aware of the existence of a larger pool of potentially eligible criteria or how large such a pool might be.

In practice, unless expert assessors consciously restrict their judgments to the criteria actually listed in the codification, they know that other criteria (or aspects) can and on occasion do emerge during the actual processes of making appraisal decisions (Sadler 2009a). Some criteria may apply to most works, others universally, but is it not uncommon

Table 3 A rubric for written responses to an assessment task

Criterion	Level A	Level B	Level C	Level D
Relevance	Consistent focus on target; all key material included; high accuracy and excellent integration	Solid level on most major issues, only moderate levels for others; some extraneous material	Generally superficial or partial; significant key aspects omitted; central problem only touched upon lightly	Overall limited or patchy inclusion of key material; many inaccuracies; main issue not addressed
Analysis	In-depth analysis; all key aspects identified, related and rigorously evaluated	Only modest abstraction beyond facts; some potential connections missed; overly derivative	Adequate in parts; gaps in logical development; many poorly supported generalisations	Almost entirely descriptive; little abstraction beyond facts; weak progression of ideas
Writing and Presentation	Exemplary in all respects, with no or very few lapses; precision evident	Technical expression sound; few errors but complex concepts not adequately conveyed	Ambiguity in stating facts; poor sentence control with frequent grammatical lapses	Poor or sloppy expression; technical aspects deficient; unclear

to make an appraisal that invokes a criterion that is simply more salient to a particular work than to others. Experienced assessors are open to those criteria that come legitimately into play only rarely and build these into their judgments when they apply. Fixed codifications do not allow for such eventualities. This reflects the reality that it may not be possible to predict with certainty the variety that can arise through students' ingenuity, originality and creativity. Obviously, even detailed elaborations do not exhaust all the properties that potentially have a bearing on determinations of quality or complex performance. Although the cross-tabulation in Table 3 may appear to be neat and tidy, latent fuzziness in its qualitative specifiers emerges as soon as it is applied. Finer and finer grained elaborations cannot fix the problem.

In the direction opposite to expansion, it is sometimes useful to conceptualise certain criteria in terms of a higher-order criterion. Still in Table 3, consistency of focus, comprehensiveness of material, accuracy, and addressing the set task (which are all stated or implied by the text cell entries) are all elements of the first main criterion, relevance. At the next level up, the three main criteria (relevance, analysis, and writing and presentation) are aspects of overall quality. At that point, the upwards process effectively stops; quality functions as the backstop criterion. In Tables 1 and 2, the backstop criterion for performance in a course or program would be competence, attainment, proficiency or achievement, all of which are high-level, integrative, abstract concepts. Ultimately, a level of performance is inferred by competent assessors who make professional judgments from evidence, which includes responses to assessment tasks and observations of behaviour. Certain skills can be practiced to the point where the level acquired can be judged in a straightforward way, but professional competence requires not only discrete knowledge and skills but also the ability to select and orchestrate them in appropriate ways to accomplish complex tasks independently, on demand and in a range of contexts (Sadler 2013a).

Now consider the quantitative specifiers. Common usage recognises that the term 'quantitative' covers amount or size generally, without being restricted to the use of numbers (for measurements, counts, indices or probability). This differs from the practice common in distinguishing quantitative research methods from qualitative methods. In statements of educational standards, quantity is generally indicated by combinations of quantifiers, modifiers and hedge words, of which there are plenty available. Embedded in Table 2 are these: excellent, strong, moderate, marginal, thorough, limited and adequate. Others are either modifiers of these or hedge words: generally, very, mostly, some, often, markedly and virtually. Which one of these is appropriate in a particular case depends on where the underlying trigger level is set. 'Moderate', for instance must have some sort of lower and upper thresholds, but who is to say how much is enough to justify the use of that term? Where is the boundary between 'limited' and 'moderate', or between 'moderate' and 'strong'? Incidentally, the mention of 'inaccurate' at the D-level presumably implies acceptable levels of accuracy for the grades above it in level. Are these to be differentiated?

The boundary problem also exists when a criterion is mentioned unaccompanied by any form of quantifier. For example skill, capacity, willingness and competence all function as both criteria *and standards* in Table 1. None of these is an all-or-nothing affair. How much skill is required for it to be said that graduates 'possess' or 'can demonstrate' it? An underlying threshold is implied, as it is with many of the listings of graduate learning outcomes, whether they carry the label 'standard' or not. As a separate issue, ability, capacity and competence each occurs in the listing of only one attribute. Is some subtle differentiation intended? And does 'skill' mean exactly the same thing in different contexts within the same degree program?

Normally, asking these sorts of questions would be regarded as pedantry, but this is a fundamental issue. The elements in progressively more detailed descriptions are intended to make meanings clearer but their specifiers are in turn of the same essence and type as those of the main elements. There is no escape. All of them are fuzzy and do not lock things down definitively. Only concrete cases and contextualisation can allow appropriate interpretations to be found, a matter taken up later. But clearly, there are 'limits as to how far the process of decomposition should proceed. While on the surface it may appear that the more detail the better, in practice there is a danger in becoming swamped with atomistic detail, at the same time losing sight of what the overall [judgment] is all about' (Sadler 1985, p. 289). Lakoff (1973) described the situation this way: '[S]tudents of language…have long been attuned to the fact that natural language concepts have vague boundaries and fuzzy edges and that, consequently, natural language sentences will very often be neither true, nor false, nor nonsensical, but rather true to a certain extent and false to a certain extent, true in certain respects and false in other respects' (p. 458). The interpretation problem is not one of granularity but of category; as mentioned above, greater specificity does not resolve it.

A typical fully specified industrial standard: eye protectors

Modern societies cannot function satisfactorily without systems of codified standards. The standards formulation selected for contrast is Australian and New Zealand Standard AS/NZS 1337.1:2010. It deals with a particular range of eye and face protectors appropriate to safety in a variety of workplaces, and the full standards specifications run to 109 pages. The interest here is in how the standards are specified, and the characteristics of the

key elements. Six of the key qualitative specifiers—the criteria—for the particular class of protectors formally designated as Medium Impact Resistance are, in summary form: Impact resistance of oculars (lenses); Transmittance and absorption of electromagnetic radiation; Thermal stability; Air flow (ventilation); Resistance to dust, splash, ignition, corrosion and other hazards; and Optical power. In the published standards, each main criterion is expanded until the detail is sufficient to purpose. Impact resistance, for instance, is expanded into: protection against shattering, penetration and fragmentation of the oculars; and distortion of the assembly that could allow contact between a projectile or debris and the eye. The criteria are mutually independent for purposes of assessing the level of protection offered. The backstop criterion is simple—safety for the eyes and face. The quantitative specifiers—the standards—set minimum performance levels for each criterion. For protectors to be stamped with the authorised symbol of the standard with which they comply prior to marketing, the evaluation rule is unequivocal: failure to meet all performance levels produces failure overall. No trading off is permitted. Only the performance specifications are set; manufacturers are therefore free to decide both the design and the materials for eye protectors.

The structure of the quantitative side of impact resistance standards is definitive, and specifies standardised testing apparatus. These non-negotiable test procedures are integral to the specification of performance levels, as the following extract shows:

> **Method for the determination of Medium Impact Resistance**
> Step (c). Project a nominally 6 mm diameter steel ball at a velocity of 45 +1.5, –0 m/s onto each of the following impact sites, repeating Steps (a), (b) and (e) for each impact site: (i) At the reference point of each ocular as given in Clause 2.4.1; (ii) At normal to the surface of the frame above the ocular, within 20 mm of the mid-line (as given in Figure 2.4) of the frame; and (iii) At 90° to straight ahead along the ocular through the centres of the front of both eyes of the headform (Appendix M, M4 Procedure, p. 79).

The conditions for passing the medium impact resistance test are equally explicit:

> **2.6.2 Performance criteria.**
> When tested in accordance with Appendix M, an ocular shall be considered to have failed: (a) if it cracks through its entire thickness into two or more pieces; (b) if more than 5 mg of the ocular material becomes detached from a part of the ocular surface remote from the surface struck by the ball; (c) if the ball passes through the ocular; or (d) if contact is made with either eye of the test headform by the ball, frame, ocular or any part or fragment of these (p. 20).

The combined specifications make reference to: concrete objects or their parts (ocular, frame, headform); materials (steel); shapes and configurations (ball, diameter, centre, midpoint, surface); physical dimensions and variables (length, time, mass, velocity, angle); standardised units measured on standardised scales (millimetres, seconds, grams, degrees); numerical measurements (5 mg, 6.00 mm, 45 m/s, 90°); and one measurement tolerance (+1.5, −0.0 m/s). None of these elements requires judgment or further interpretation; none foreshadows that finer grained specifications could be helpful. For all practical purposes, the possibility of hierarchical decomposition that is possible with achievement standards can go no further for impact resistance as a property. It simply stops.

The sections of the AS/NZS 1337 standards specifications that apply to various levels of eye protection, including Medium Impact Resistance protectors, are arbitrary in that they were decided by human agency following a set of procedures laid down in 2010 jointly by

Standards Australia and Standards New Zealand, the authorities which published the standards. These procedures employ consultative processes involving wide representation from experts in workplace health and safety, manufacturers of safety equipment, end users and the public at large. They constitute, therefore, true standards. Once settled, the standards specifications remain fixed until the next round of formal revision. (The previous version was published in 1992.) To the extent that the constituent terms require interpretation, the interpretation is made in a direct and secure way—the specifications can be taken literally and applied by any competent person who understands the text, formulae and diagrams; has access to appropriate materials and testing equipment; and possesses the know-how to use them all properly. The person must know, for instance, the meanings of terms—steel, ball, 6 mm, diameter, velocity, 45 m/s, and so on. However, those meanings are identical within any scientifically literate community and would be interpreted and applied in exactly the same way. They are definitive and intrinsically valid, with no matters left to individual human judgment at the point of implementation. Furthermore, the various criteria are separable in concept and practice. Test rigs can be built on demand at any time or place. Provided all the rules are followed exactly, various instantiations of the testing are equivalent in composition, functionality and rigour—and produce the same results.

These types of standards possess their authority not only through the organisations which issue them and monitor their use but also through the universality and precision of their formulation. The relative aspect of word-quantifiers in education signified by such words as high, moderate and low does not arise. A 6 mm steel ball is not large, medium or small; it is simply 6 mm in diameter. Of course, whether Medium Impact Resistance is adequate for a particular workplace remains a matter for case-by-case judgment, but that issue does not form part of the standards specifications.

Further analysis of achievement standards codifications

Qualifiers, modifiers and hedge words in educational standards statements are typically interpreted relatively rather than absolutely. For rubrics, their meanings may be limited to a cohort. This applies to both criteria as properties or features (such as accuracy, integration, logic and precision) and quantitative specifiers (high, some, poor, frequent). Superlatives in ordinary language are relativistic. A work that is 'outstanding' literally stands out from some background, either real and immediate, or recalled from memory. The same applies to extraordinary, exceptional, excellent and superb. In some situations, the contextuality is made explicit, as when an awarded grade is tempered with such phrases as 'for this cohort', 'for the types of students enrolled in this program', 'given the disruptions this semester', or 'for studies terminating at the bachelor level'. In general, the specifiers used in achievement standard codifications statements are highly elastic. Their interpretations readily expand or contract so as to 'fit' or cover a particular reality, with perhaps some modification by residual knowledge about other contexts, especially those previously encountered personally by a marker.

Cohort-based 'standards' exploit the elasticity of the terms in the codification. A practice recommended in many books on assessment is to begin marking or grading a batch of student works by first establishing a baseline. This involves scanning and reviewing a 'range sample' of student works to get a feel for the broad level of performance. On a larger scale, a higher education institution may deliberately take into account the social backgrounds and entry levels of its students and interpret the outcome specifications accordingly. Such practices openly legitimate cohort-based existential 'standards' which

are incompatible with the concept of true standards. Elasticity in the terms also explains why a particular codification (such as a rubric) can be found to apply equally well to first year undergraduate, final year and graduate level student works. The codification does not change; the meanings of the specifiers are reinterpreted to accommodate whatever level of student work they are applied to. Although that reflects the power and flexibility of language generally, it is of little value when trying to anchor and convey standards through codification alone. Research on the elasticity of meanings of common terms has been the subject of research for over 50 years (Helson et al. 1956; Helson 1959), and is an element of the wider phenomenon of human adaptability which pervades human lives and the making of meaning.

Does supplementing codifications *with exemplars* supply the necessary anchorage? The combination of a codification and its 'associated' exemplars requires a two-stage interpretation. As a set of words, a codification is not, in general, capable of describing a particular exemplar exactly. It cannot function as the 'decider' as to whether or not another instance qualifies as a member of the class. By their very nature, codifications of educational standards are generic descriptions of classes of things. For some works, aspects which are crucial in making correct determinations of quality or level may not appear in the codification at all. Works which competent appraisers regard as of equivalent quality or level typically differ from one another, sometimes in significant ways. In other words, people can and do regularly make judgments about different objects and classify them in the same 'worth-class'. Wittgenstein (1953) labelled this as noticing 'family resemblances' that are sufficiently pronounced for things to be safely classified; Dreyfus and Dreyfus (1986) called it 'holistic similarity recognition' (p. 28). Despite substantial differences in content or form, different student works are regularly judged to be equivalent in value, worth or quality. Competent assessors notice or perceive different patterns of characteristics (cues), which they are often able to identify explicitly only during or after a judgment is made. Contrary to the assumptions underlying the principles of codification, this is by no means an insurmountable problem. Indeed, '[t]he process by which equivalent judgments can result from different patterns of cues is central to any theory of judgment' (Einhorn et al. 1979).

Finally, the process of codification places a specific interposition between the primary evidence of achievement and the grading judgment. This naturally puts the focus of attention on the content and structure of the codification (which is concrete) rather on than the underlying standard it is supposed to represent (which is abstract). Abercrombie's (1969) work on influences which bias human perception is salutatory: 'How to tell students what to look for without telling them what to see is the dilemma of teaching'. It is also the dilemma of making appraisals. A codification is inherently constraining, even more so when accompanied by 'exemplars', because its formulation does not lend itself to the admission of new types and configurations of qualities that constitute overall quality. Although codification may purport to increase the efficiency of informational exchange, even a fully effective formulation of [codification + exemplars] would come at a potentially significant cost—rigidity and uniformity. 'The need for [codifications that are] mutually understandable within the organization imposes a uniformity requirement on the behavior of the participants. They [become] specialized in the information capable of being transmitted by the [codifications], so that … they learn more in the direction of their activity and become less efficient in acquiring and transmitting information not easily fitted in the [codification]. Hence, the organization itself serves to mold the behavior of its members' (Arrow 1974, p. 56–57). (Arrow's original term 'code' has been replaced here by 'codification', the term most commonly used in current research literature.) What is

required is to hold firm on the standards to be applied but be quite open to different ways of seeing the underlying standard expressed.

A possible way forward

The object of the exercise is clear enough—how to facilitate trustworthy discriminations among levels of academic achievement in a devolved system—but the analysis above, cast as it is in the form of a contrast between standards which are and which are not fully specifiable, points to the futility of codification as an approach, despite widespread commitment to it. In their critical analyses of modern trends towards codification in assessment criteria, standards and quality assurance, González-Arnal and Burwood (2003) concluded that the assumption of the positive benefits of codification in higher education is deeply—and for the most part uncritically—embedded in academic culture. This section is an attempt at resolving the apparent dilemmas in moving forward on standards by responding to three questions. If codifications fail to convey standards, are 'standards' needed at all? If so, is some form of external representation of them, a material form, necessary? ('Material form' is a term drawn from copyright law. It refers to any form of information storage that is sufficiently permanent or stable for the information to be perceived, identified, reproduced and communicated on demand.) The final question is this: If codification with or without exemplars is not the way ahead, what other approach could potentially serve to anchor achievement standards for future use?

Early in this paper, a distinction was made between criteria and standards, but that conceptualisation of standards did not go far enough. 'Standard' turns out to be a troublesome concept in higher education because different meanings can be implied even in the same seemingly straightforward discussion. At this point, the meaning of a standard is given a bit more formal substance. This is necessary for developing answers to the three questions. A standard, then, is to be taken as a 'definite degree of academic achievement established by authority, custom, or consensus and used as a fixed reference point for reporting a student's level of attainment' (Sadler 2013b). This is consistent with the usage for eye protector standards, and equally for a wide range of other standards in society. The standards are performance or reference levels set by a recognised authority as a deliberate act. As and when necessary, such standards may be reset, also as a deliberate act.

To say 'standards are falling' to mean that student performance levels are steadily decreasing uses the term inappropriately because it interprets standards as empirically determined average levels of performance. This is incompatible with standards as fixed reference levels. If standards could be fixed and held stable over time, student achievement could be graded with integrity (Sadler 2009b), the performance of different cohorts of students could be compared, research on the effectiveness of teaching could be carried out, and general achievement levels in an academic program or institution could be mapped and evaluated longitudinally. This would therefore go a long way towards addressing the quality assurance question. The need for standards as fixed reference levels is therefore crucial.

Where teachers create codifications in the form of rubrics or criteria-standards specifications, the ostensible purposes are to increase marking transparency and to guide appraisers in their marking. (That said, teachers may not establish the full meaning and implications of even their own guidelines until they have carried out a baseline survey as described above.) Particularistic 'standards' that are set for and within courses do not satisfy the above definition because they are disconnected from similarly devised

'standards' in other courses. Suppose that a group of assessors in different courses collaborate to moderate standards among themselves, perhaps even across degree programs and institutions. Even these are not true standards, because they lack anchorage; one group's 'standards' may well differ from another group's 'standards'. Besides, there would still be no way of detecting whether the 'standards' are drifting over time. Comparability across both contexts *and time* requires that standards are not only held in common, but also kept secure until the need arises to revise them.

Moving to the second question, must what is to be made secure be expressed in material form? The extensive literature on the limitations on human information processing indicates that it is. Assessors in general find it difficult to hold standards constant in their heads in any sort of absolute way that allows them to be applied on demand (Stewart et al. 2005). Without anchorage, even 'memorised' reference points for judgments adapt to the circumstances. However, people generally can retain the ability to compare things and make consistent discriminations among them, a finding which is well documented. They are just not good at holding absolutes unaided. It turns out that judgments among relevant objects in new settings can be made reliably provided the objects are considered two at a time, to facilitate pair-wise or 'paired' comparisons (Thurstone 1927; Saaty 1977). If standards can be expressed in a material form which is similar enough to the objects being appraised for their relative quality to be determined, (new) works can be put up against (fixed) standards, and the anchorage issue is resolved. This would permit competent assessors to make consistent standards-referenced judgments of even single works at arbitrary times and places. If all this sounds highly theoretical, the proposed solution is relatively straightforward.

An alternative to codification

Although the qualitative and quantitative terms used in codifications of achievement standards lack the necessary linguistic properties to carry true standards successfully, words are the very stuff of explanations. They can play an indispensable part in assembling a material form useful for conveying an achievement standard in a way that is consistent with how complex qualitative judgments are made. The material form is not of the underlying standard itself, which is an abstraction, but of the effect of applying the standard, referred to here as an *instantiation* (of a judgment against the standard). This is similar in principle to an approach suggested by Popham (1994) in a somewhat related context. A way to do this for a single student's grade in a single course is to first identify a body of student work that is judged, for example, as definitely worth a Pass—clearly not a Fail but not up to Merit level. The level of achievement is inferred from this body of student work. Such an instantiation of a judgment is indicative of the underlying standard of Pass. (In a standard-setting situation, this decision would need to be made after appropriate deliberation among competent judges.) This implies sensory input which results in an evaluative conclusion about a concrete object or observed behaviour.

When agreement is reached, the next step would be to explain exactly why such a classification is warranted. Such an explanation would draw attention to those aspects of the work, whether positive or negative, that support the decision, and ideally would make use of at least some of the 'graduate outcomes' listed in the Introduction to this article for reasons explained in (Sadler 2013b). Such an explanation, if properly constructed, means that the text would function as an evaluative description of an actual concrete object. If the text is an adequate account of the evaluative judgment, the meaning of its specifiers will be fully contextualised and clear. The specifiers in the text then have a true *referent* (in the

semiotic sense). The match between the two is what matters, because all the qualitative and quantitative specifiers have their interpretations fixed or grounded in reality.

The language used to validate the judgment need not be drawn from a standard vocabulary if other words explain it better; the assessor has complete control over how the explanation is structured and phrased and also the granularity of what is reported. If how the work comes together as a whole is particularly noteworthy but not in terms of any easily identifiable characteristics, the justification may find such terms as flair, flow or artistry more appropriate than any others. The fact that the meanings of these terms are ordinarily hard to pin down is countered by the particularity of the referent to which they specifically refer. As another example, a particular work may show an exceedingly high level of mastery over the task but achieve this by quite unorthodox means, such as deliberately breaking some rule or ignoring some convention expressly for the strength this gives the work as a whole. Attention would need to be drawn to this.

The features of the object and the text of the justification for the assigned code therefore stand in a mutually reciprocal relationship. The aim in constructing the text is to give conceptual substance to the qualitative specifiers and reduce the elasticity of the quantitative specifiers. Aspects that are latent for the valuation do not need to have attention drawn to them. The set of criteria would be deliberately left open to allow the description to be fully responsive to the properties of the student work. This is a far cry from the emphasis on characteristic features around which codifications are drawn up, and also from properties held in common. Perception and professional judgments clearly play fundamental roles in achievement standards systems, and both need to be tuned. On the perception front, an aspect is considered worthy of mention in a valuation when two conditions are satisfied. The first is that the identification of the aspect has clearly passed some perception threshold (which is in essence quantitative). The second is that the aspect has an intimate connection with the overall judgment, regardless of whether its identification was part of deliberative judgment making—or simply followed the final decision reflectively. Both conditions are involved in appraising work through evaluative eyes.

Clearly, a standard as an abstraction cannot be adequately inferred from a single instantiation. What is needed is a range of instantiations for, say, each grade band, each instantiation consisting of a triplet composed of actual student work, the assigned code (mark or grade) and the justification. Each instantiation offers a viewing point (as it were) on the (abstract) standard, the aim being to fix the standard's position in the evaluative decision space (to invoke the language of satellite-based positioning systems). The number of points to be fixed has to be enough to fix an independent baseline for assigning the appropriate code to each judgment. Triplets provide both the mechanism and the critical insights into the standards being applied, and have potential for accelerating the process of experiential learning as compared with unguided induction purely from extensive sets of examples.

If this conjures up the prospect of a mountainous archive of triplets, a few comments may help to allay fears. First, a critical interface in any standards system is the one between satisfactory and unsatisfactory, pass and fail. The greatest return on investment is probably to be had by focusing on that interface. This is what 'benchmark', 'threshold', or 'minimally competent' decisions are essentially about. The second most critical point is probably the level of attainment necessary to warrant award of the highest available code. Intermediate codings can be interpolated by capable assessors, and should be left with them as part of their professional academic responsibility.

In practice, several refinements would be necessary for such a system to work. The above description of triplets has referred only to student work. By itself, student work

does not constitute the full evidence required for a judgment of achievement. An important consideration is the structure and quality of the assessment task—its design and its specifications. An assessment task which is deficient in its design or specifications cannot be relied upon to elicit valid student responses. Poor quality evidence of a student's level of achievement must not be confused with evidence of poor achievement. To that must be added the conditions under which responses were produced. All this so far should not be taken as sufficient on its own as a way of fixing standards in an assessment system characterised by distributed responsibility for making judgments about student achievement. An investment needs to be made in educating assessors into the overall strategy, especially its rationale. Just as the terms in industrial standards codifications assume a scientifically literate community, so the development of corresponding standards literacy is necessary in higher education. Consideration of these topics, which are significant in their own right, lies outside the scope of this article, but details of some approaches to them are outlined in two companion articles (Sadler 2011, 2013b).

This proposal for an alternative material form has a clear precedent in principle in the approach taken by biological taxonomists. They had long sought tight codifications for a biological classification system for organisms, but the complexity of life forms seemed to defy all attempts at definitive classification. The history of taxonomic development, including several myths that received wide currency only later shown to be incorrect, attempts at formulating biological divisions by means of common features, the subsequent use of selected (exemplary) specimens accompanied by detailed descriptions, and the place of the relevant literacy among competent practitioners (including the centrality of professional judgments) have been systematically documented by Winsor (2003). In short, human judgment and the acceptance of family resemblances rather than an insistence on sharply defined boundaries between divisions have remained essential ingredients in taxonomic processes all the way from Linnaeus onward.

Contextual and historical note

The issue of codification and specifiability is as ancient as Socrates (Dreyfus and Dreyfus 1984). Modern research on the topic has been published in fields as diverse as agriculture, artificial intelligence, business, control theory, economics, engineering, higher education, history and philosophy of science, linguistics, management, medicine, microbiology, organisation theory, philosophy, psychiatry, psychology, sociology, social anthropology and taxonomy. Despite differences in terminology and reasoning, strong elements of convergence have been identified (Needham 1975; Winsor 2003). Knowledge transfer or dissemination by means of descriptive statements stored in appropriate media is theoretically possible only under certain conditions. The practical problem is deciding whether knowledge of a certain type satisfies the conditions.

In this article, the analysis has dealt primarily with the fuzziness and elasticity of specifiers. Only touched upon briefly is a complementary line of analysis which leads to the same conclusion. It emphasises specifieds rather than specifiers, in particular, specifieds related to certain classificatory systems. When membership of a class is based on a distinctive set of attributes which all class members have in common, codification of the rules for class membership is theoretically possible. However, codification is difficult or impossible when this condition is not met. Such classes abound. Some are concepts which are part of everyday life and discourse, among them 'fairness', 'game' and 'furniture'.

Wittgenstein (1953) puzzled repeatedly over what makes a game a game. Others are fundamental to the development and use of taxonomies in the biological sciences. What makes orderly classification feasible is that people with appropriate experience are readily able to recognise similarity, even when it is complex. The same applies to academic standards. It would be entirely unnecessary to compare every student's work with the standard. Competent assessors would make most judgments without reference to standards documentation, but at any time, the standards anchoring framework is there for periodic reference in checking judgments.

Obviously, not even a brief excursion through this rich literature is possible here. Suffice to say that the references included in this article are highly selective, being those most directly pertinent to the development of the argument and conclusion.

Conclusion

Assuring standards in higher education is a major concern in many countries. Typical agenda include teaching quality, student satisfaction, student services, resources, spaces, and more recently, academic achievement standards. For the last of these, the challenge has been to devise a strategy whereby judgments about the comparability of different levels of underlying academic achievement can be made by different judges, in different places, at different times from different evidence. That requires a clear understanding about what 'underlying achievement' is, and how it can be manifested and recognised. One particular approach to the last of these forms the focus for this article.

Projects for developing codified 'standards' have engaged in broad consultative processes, the expectation being that the resulting explicit specifications will communicate shared standards and lead to improved integrity in grading. This in turn should enable higher education institutions and directors of academic programs and courses to be held more accountable for the quality of academic achievement and graduate learning outcomes. Discipline-based 'standards' codifications can certainly function as valuable tools for guiding curriculum planning, academic program development, teaching and assessment. Furthermore, consultative processes engaged in during their development can promote consensus and commitment to the concept of shared standards as academic values, and to the pursuit of higher-order outcomes integrated with discipline content. But that is where the benefits of codification stop. They cannot safeguard academic achievement standards or lead to high levels of comparability in judgments of student performance.

The key operational elements in codifications of standards generally are criteria (or qualities) and specified minimum levels on those qualities. A host of codified standards in industry and commerce can achieve their purpose because their key elements are unambiguous and independent of context. However, the analysis in this article has shown that the key elements in codifications of academic achievement standards lack the appropriate linguistic properties. They are inherently fuzzy and open to interpretation, and the minimum required levels are invariably expressed in relative rather than absolute terms. This means that although the overall structures of the two categories of standards may on the surface appear to be similar, achievement standards codifications cannot deliver on what is expected of them. A new approach outlined in this article would make use of a different material form. This, when complemented by appropriate moderation and calibration processes described elsewhere, holds out improved prospects for assuring academic achievement standards.

References

Abercrombie, M. L. J. (1969). *The anatomy of judgement: An investigation into the processes of perception and reasoning*. Harmondsworth, Middlesex: Penguin.

Amin, A., & Cohendet, P. (2004). *Architectures of knowledge: Firms, capabilities and communities*. Oxford, NY: Oxford University Press.

Arrow, K. J. (1974). *The limits of organization*. New York: Norton.

Dreyfus, H. L., & Dreyfus, S. E. (1984). From Socrates to expert systems: The limits of calculative rationality. *Technology in Society, 6*(3), 217–233.

Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine: The power of human intuition and expertise in the era of the computer*. Oxford: Basil Blackwell.

Einhorn, H. J., Kleinmuntz, D. N., & Kleinmuntz, B. (1979). Linear regression and process-tracing models of judgment. *Psychological Review, 86*(5), 465–485.

González–Arnal, S., & Burwood, S. (2003). Tacit knowledge and public accounts. *Journal of Philosophy of Education, 37*(3), 377–391.

Helson, H. (1959). Adaptation level theory. In S. Koch (Ed.). *Psychology: A study of a science.* vol 1, *Sensory, perceptual, and physiological formulations.* New York: McGraw-Hill.

Helson, H., Dworkin, R. S., & Michels, W. C. (1956). Quantitative denotations of common terms as a function of background. *The American Journal of Psychology, 69*(2), 194–208.

Lakoff, G. (1973). Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic, 2*(4), 458–508.

Needham, R. (1975). Polythetic classification: Convergence and consequences. *Man (New Series), 10*(3), 349–369.

Popham, W. J. (1994). The instructional consequences of criterion-referenced clarity. *Educational Measurement: Issues and Practice, 13*(4), 15–18–30.

Saaty, T. L. (1977). A scaling method for priorities in hierarchical structures. *Journal of Mathematical Psychology, 15*(3), 234–281.

Sadler, D. R. (1985). The origins and functions of evaluative criteria. *Educational Theory, 35*(3), 285–297.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education, 13*(2), 191–209.

Sadler, D. R. (2009a). Indeterminacy in the use of preset criteria for assessment and grading in higher education. *Assessment and Evaluation in Higher Education, 34*(2), 159–179.

Sadler, D. R. (2009b). Grade integrity and the representation of academic achievement. *Studies in Higher Education, 34*(7), 807–826.

Sadler, D. R. (2011). Academic freedom, achievement standards and professional identity. *Quality in Higher Education, 17*(1), 103–118.

Sadler, D. R. (2013a). Making competent judgments of competence. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn, & J. Fege (Eds.), *Modeling and measuring competencies in higher education: Tasks and challenges* (pp. 13–27). Rotterdam: Sense Publishers.

Sadler, D. R. (2013b). Assuring academic achievement standards: From moderation to calibration. *Assessment in Education: Principles, Policy and Practice, 20*(1), 5–19.

Standards Australia and Standards New Zealand. (2010). *AS/NZS 1337.1:2010 Australia/New Zealand Standard. Personal eye protection.* Part 1: Eye and face protectors for occupational applications. Sydney: Standards Australia and Wellington: Standards New Zealand.

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review, 112*(4), 881–911.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273–286.

Winsor, M. P. (2003). Non-essentialist methods in pre-Darwinian taxonomy. *Biology and Philosophy, 18*(3), 387–400.

Wittgenstein, L. (1953). *Philosophical investigations*. (trans.) G. E. M. Anscombe. Oxford: Blackwell.